BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images may include (but are not limited to):

- BLACK BORDERS
- TEXT CUT OFF AT TOP, BOTTOM OR SIDES
- FADED TEXT
- ILLEGIBLE TEXT
- SKEWED/SLANTED IMAGES
- COLORED PHOTOS
- BLACK OR VERY BLACK AND WHITE DARK PHOTOS
- GRAY SCALE DOCUMENTS

IMAGES ARE BEST AVAILABLE COPY.

As rescanning documents *will not* correct images, please do not report the images to the Image Problem Mailbox.

(19)日本国特許庁（ＪＰ）　(12)公開特許公報（A）　(11)特許出願公開番号

特開平6－215184

(43)公開日　平成６年(1994)８月５日

(51)Int.Cl.[5]	識別記号	庁内整理番号	ＦＩ	技術表示箇所
Ｇ０６Ｋ　9/36				
Ｇ０６Ｆ　15/70	３３０　Ａ	9071－5L		
Ｇ０６Ｋ　9/20	３４０　Ｋ			
9/72		9289－5L		

審査請求　未請求　請求項の数３　ＦＤ　（全 10 頁）

(21)出願番号　特願平5－249966

(22)出願日　平成５年(1993)９月10日

(31)優先権主張番号　特願平4－273784

(32)優先日　平４(1992)９月17日

(33)優先権主張国　日本（ＪＰ）

(71)出願人　000237156
富士ファコム制御株式会社
東京都日野市富士町１番地

(71)出願人　000005234
富士電機株式会社
神奈川県川崎市川崎区田辺新田１番１号

(72)発明者　森　泰二
東京都日野市富士町１番地　富士ファコム制御株式会社内

(72)発明者　本郷　保夫
東京都日野市富士町１番地　富士ファコム制御株式会社内

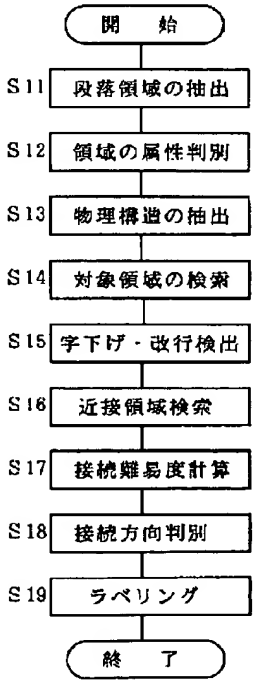
(74)代理人　弁理士　森田　雄一

(54)【発明の名称】　抽出領域のラベリング装置

(57)【要約】

【目的】　文書画像から抽出した段落領域のラベリング精度を向上する。

【構成】　最初に、入力された文書画像から段落領域を抽出するとともに（Ｓ１１）、抽出された領域の属性を判別し（Ｓ１２）、属性が文字である段落領域についての物理構造を解析・抽出して探索木を作成する（Ｓ１３）。次に、属性が見出しである段落領域を挟む位置の文書段落領域を検索して対象領域とし（Ｓ１４）、抽出された各対象領域について、字下げ・改行の有無を検索する（Ｓ１５）。また、各段落領域が相互に接続する可能性のある近接領域を検索し（Ｓ１６）、さらに、各段落領域間の接続難易度をそれぞれ計算することにより（Ｓ１７）、対象の文書段落領域の接続方向を判別する（Ｓ１８）。得られた接続方向に基づき各文書段落領域の接続を行い、抽出領域をラベリングする（Ｓ１９）。



1

【特許請求の範囲】

【請求項１】　文書画像から抽出された段落領域の属性が文書であるか否かを判別する手段と、

属性が文書であると判別された文書段落領域から探索木を作成する手段と、

属性が文書でないと判別された非文書段落領域の前後に位置しかつ他の文書段落領域との接続が可能な文書段落領域の組合わせを探索木から検索する手段と、

検索された文書段落領域ごとに最終行の行末空白および先頭行の行頭空白を検出する手段と、

文書段落領域の先頭および末尾の空白の有無から接続の組合わせごとに文書段落領域間の接続の難易度を算出する手段と、

文書段落領域間の接続の組合わせごとの接続難易度を比較して非文書段落領域前後の文書段落領域の接続方向を判別する手段と、

判別された接続方向に基づき各文書段落領域のラベリングを行う手段と、

を備えたことを特徴とする抽出領域のラベリング装置。

【請求項２】　文書画像から抽出された段落領域の属性が文書であるか否かを判別する手段と、

属性が文書であると判別された文書段落領域から探索木を作成する手段と、

属性が文書でないと判別された非文書段落領域の前後に位置しかつ他の文書段落領域との接続が可能な文書段落領域の組合わせを探索木から検索する手段と、

検索された文書段落領域ごとに文字認識を行う手段と、

文書段落領域の先頭および末尾の認識結果を接続の組み合わせごとに比較して両者の接続の適不適から文書段落領域間の接続の難易度を算出する手段と、

文書段落領域間の接続の組合わせごとの接続難易度を比較して非文書段落領域前後の文書段落領域の接続方向を判別する手段と、

判別された接続方向に基づき各文書段落領域のラベリングを行う手段と、

を備えたことを特徴とする抽出領域のラベリング装置。

【請求項３】　請求項１記載の抽出領域のラベリング装置において、請求項２記載の抽出領域のラベリング装置における文字認識手段および接続難易度算出手段を備え、文書段落領域間の接続の組合わせごとの接続難易度を接続方向別に比較し、接続難易度の値の差が所定値以下であれば、請求項２記載の文字認識手段および接続難易度算出手段により接続難易度を求めることを特徴とする抽出領域のラベリング装置。

【発明の詳細な説明】

【０００１】

【産業上の利用分野】本発明は、文字認識装置に入力された文書画像から文字領域を抽出した後に、抽出領域をラベリングする抽出領域のラベリング装置に関する。

【０００２】

2

【従来の技術】従来の文字認識装置における文字領域の抽出およびラベリングは、図１２に示すフローチャートのように行われている。図では、最初に文書画像が２値画像データとして入力されると（Ｓ２１）、ぼかし等の処理を施した後、その周辺部のゴミ等を除去し（Ｓ２２）、段組構造を判定する（Ｓ２３）。次いで、縦書き・横書きを判定してから（Ｓ２４）、行ブロックの抽出および行ブロックのラベリングを行う（Ｓ２５，２６）。さらに、段ブロックの抽出をした後に（Ｓ２７）、段ブロックのラベリングを行う（Ｓ２８）。この段ブロックのラベリングでは、段落の物理的な位置関係から探索木を作成し、予め設定されたルールに基づいて枝を検索して段落の接続順を決定していた。

【０００３】

【発明が解決しようとする課題】しかしながら、このような段落のラベリング方法では、新聞等の複雑な構造を有する文書に対して、必ずしも妥当なラベリングが行われるとは限らない。例えば、図１３に示すような新聞記事の場合、見出し⑤が文字領域を区分するセパレータとして認識されてしまうと、その前後の段落③，⑥および④、⑦は見出し⑤を飛び越して接続されることがなく、誤ってラベリングされてしまう。同様に、図１４に示すような新聞記事の場合、見出し⑥がセパレータとして認識されてしまうと、その前後の段落③，⑦および④、⑧は見出し⑥を飛び越して接続されることがなく、誤ってラベリングされてしまうという問題があった。本発明は上記問題点を解決するためになされたもので、その目的とするところは、見出しをセパレータとして認識することなく常に正確にラベリングすることのできる抽出領域のラベリング装置を提供することにある。

【０００４】

【課題を解決するための手段】上記目的を達成するために、第１の発明は、文書画像から抽出された段落領域の属性が文書であるか否かを判別する手段と、属性が文書であると判別された文書段落領域から探索木を作成する手段と、属性が文書でないと判別された非文書段落領域の前後に位置しかつ他の文書段落領域との接続が可能な文書段落領域の組合わせを探索木から検索する手段と、検索された文書段落領域ごとに最終行の行末空白および先頭行の行頭空白を検出する手段と、文書段落領域の先頭および末尾の空白の有無から接続の組合わせごとに文書段落領域間の接続の難易度を算出する手段と、文書段落領域間の接続の組合わせごとの接続難易度を比較して非文書段落領域前後の文書段落領域の接続方向を判別する手段と、判別された接続方向に基づき各文書段落領域のラベリングを行う手段とを備えたことを特徴とする。

【０００５】第２の発明は、文書画像から抽出された段落領域の属性が文書であるか否かを判別する手段と、属性が文書であると判別された文書段落領域から探索木を作成する手段と、属性が文書でないと判別された非文書

段落領域の前後に位置しかつ他の文書段落領域との接続が可能な文書段落領域の組合わせを探索木から検索する手段と、検索された文書段落領域ごとに文字認識を行う手段と、文書段落領域の先頭および末尾の認識結果を接続の組み合わせごとに比較して両者の接続の適不適から文書段落領域間の接続の難易度を算出する手段と、文書段落領域間の接続の組合わせごとの接続難易度を比較して非文書段落領域前後の文書段落領域の接続方向を判別する手段と、判別された接続方向に基づき各文書段落領域のラベリングを行う手段とを備えたことを特徴とする。

【０００６】第３の発明は、第１の発明に、第２の発明における文字認識手段および接続難易度算出手段を備え、文書段落領域間の接続の組合わせごとの接続難易度を接続方向別に比較し、接続難易度の値の差が所定値以下であれば、第２の発明の文字認識手段および接続難易度算出手段により接続難易度を求めるようにしたことを特徴とする。

【０００７】

【作用】第１の発明においては、文書画像から抽出された段落領域の属性が文書であるか否かが判別され、属性が文書であると判別された文書段落領域から探索木が作成される。属性が文書でないと判別された非文書段落領域の前後に位置しかつ他の文書段落領域との接続が可能な文書段落領域の組合わせが探索木から検索される。次いで、検索された文書段落領域ごとに最終行の行末空白および先頭行の行頭空白が検出され、空白の有無から接続の組合わせごとに文書段落領域間の接続の難易度が算出される。さらに、算出された接続難易度が比較されて非文書段落領域前後の文書段落領域の接続方向が判別され、その接続方向に基づき各文書段落領域のラベリングが行われる。

【０００８】第２の発明においては、文書画像から抽出された段落領域の属性が文書であるか否かが判別され、属性が文書であると判別された文書段落領域から探索木が作成される。属性が文書でないと判別された非文書段落領域の前後に位置しかつ他の文書段落領域との接続が可能な文書段落領域の組合わせが探索木から検索される。次いで、検索された文書段落領域ごとに文字認識が行われ、文書段落領域の先頭および末尾の認識結果が接続の組み合わせごとに比較され、両者の接続の適不適から文書段落領域間の接続の難易度が算出される。さらに、算出された接続難易度が比較されて非文書段落領域前後の文書段落領域の接続方向が判別され、その接続方向に基づき各文書段落領域のラベリングが行われる。

【０００９】第３の発明においては、第１の発明に、第２の発明における文字認識手段および接続難易度算出手段が備えられており、文書段落領域間の接続の組合わせごとの接続難易度が接続方向別に比較され、接続難易度の値の差が所定値以下であれば、第２の発明の文字認識

手段および接続難易度算出手段により接続難易度が求められる。

【００１０】

【実施例】以下、図に沿って本発明の実施例を説明する。図１は第１の発明の実施例の処理動作を示すフローチャートである。図において、最初に、入力された文書画像から段落領域を抽出するとともに（Ｓ１１）、抽出された領域の属性を判別する（Ｓ１２）。ここで判別される属性としては、文書、見出し、図、写真等がある。次いで、属性が文書である段落領域についての物理構造を解析・抽出する（Ｓ１３）。ここで抽出された物理構造より探索木が作成される。

【００１１】さらに、属性が文書ではなく見出し等である段落領域を挟む位置の文書段落領域を検索して対象領域とする（Ｓ１４）。このように探索木を検索すると、例えば図２に示されるように、見出し領域１にかかる２段に、それぞれ前後して位置する文書段落領域２～５が抽出される。次に、抽出された各対象領域について、字下げ・改行の有無を検索する（Ｓ１５）。具体的には、内蔵する行の座標や投影等の画像処理により、段落領域先頭の字下げと段落領域最後の改行による空白を検出してマークする。また、各段落領域が接続する可能性のある近接領域を検索する（Ｓ１６）。この近接領域の検索は図３のように行われる。

【００１２】図では、見出し領域６の前の段落領域７が、見出し領域６を飛び越す場合に文書段落領域８へ接続され、飛び越さない場合に段落領域９へ接続される。同様にして、見出し領域６の後の段落領域８の接続先についても、文書段落領域９へ接続される場合と文書段落領域１０へ接続される場合がある。つまり、見出し領域６がセパレータとして機能する場合は、文書段落領域７～１２等の接続方向がそれぞれ縦方向となるが、見出し領域６がセパレータとして機能しない場合の接続方向はそれぞれ横方向となる。

【００１３】次に、各段落領域間の接続の組み合わせごとに接続難易度を計算する（Ｓ１７）。具体的には図４に示すように、前の文書段落領域が改行で終了し後の文書段落領域が字下げで始まっている場合は接続の可能性が大であるから接続難易度を＋１とし、前の文書段落領域が改行で終了し後の文書段落領域に字下げがない場合は接続の可能性が小であるから接続難易度を－１とする。また、前の文書段落領域に改行がなく、後の文書段落領域に字下げがある場合は接続の可能性が小であるから接続難易度を－１とする。

【００１４】さらに、両文書段落領域とも空白がない場合は接続関係が不明であるものとして接続難易度を０とする。また、これ以外にも、両文書段落領域に空白がなくしかもその間に見出しが位置している場合は、両者が接続される可能性が極めて高いものとして接続難易度を１．５とする。その理由は、見出しがセパレータである

5

なら後の文書段落領域にも必ず字下げがあるはずだからである。さらに、上述した以外にも、文書段落領域の配置上のルールにより接続難易度を設定することが可能である。

【００１５】次いで、得られた接続難易度を対象の文書段落領域の中の見出しの前または後における縦方向、および見出しをまたぐ横方向ごとにそれぞれ累積し、その縦と横の累積値を比較することにより対象の文書段落領域の接続方向を判別する（Ｓ１８）。その結果、得られた接続方向に基づいて、各文書段落領域の接続を行い、抽出領域のラベリングを終了する（Ｓ１９）。上述した実施例の方法で図１３の文書段落領域をラベリングした結果が図５となる。この第１の発明の実施例では、文書段落領域の間に見出しがある場合でも、その前後の文書段落領域の空白の有無により接続方向が比較的高速に判定されて正しくラベリングが行われるようになる。

【００１６】次に、第２の発明の実施例について説明する。図６は第２の発明の実施例の処理動作を示すフローチャートである。図において、最初に、入力された文書画像から段落領域を抽出するとともに（Ｓ３１）、抽出された領域の属性を判別する（Ｓ３２）。ここで判別される属性としては、文書、見出し、図、写真等がある。次いで、属性が文書である段落領域についての物理構造を解析・抽出して、探索木を作成する（Ｓ３３）。図７は、図１４に示す文書画像から作成された探索木を示す。図中のＨは文書段落領域であるところの本文領域を、Ｍは見出し領域を表す。

【００１７】さらに、属性が文書である領域について順次文字認識を行い（Ｓ３４）、得られた文字コードを段落領域情報等とともにメモリに記憶する（Ｓ３５）。次に、探索木より、図、写真、見出し等の本文以外の領域を検索し、着目領域とする（Ｓ３６）。さらに着目領域が本文領域に挟まれているか否かによりセパレータであるか否かを判別し（Ｓ３７）、挟まれていなければ着目領域がセパレータであるものとしてＳ４３へ進みラベリングを行う（Ｓ３８Ｙｅｓ）。

【００１８】挟まれていれば（Ｓ３８Ｎｏ）、着目領域に隣接する本文領域を探索木より検索し（Ｓ３９）、さらに検索された本文領域相互間における接続の可能性を検索する（Ｓ４０）。これらの近接領域の検索は第１の発明の実施例と同様に図３のように行われる。次に、検索により得られた本文領域間の接続難易度を求める（Ｓ４１）。ここで求められる接続難易度とは、Ｓ３４において領域ごとに認識された文章を言語処理技術を用いて単語に分割し、次いで、段落間をまたぐ単語について予め用意した単語辞書との照合結果および文法ルールによる文法チェックの結果を接続確率等のように数値化し接続難易度としたものである。図８は、これらの接続難易度計算の処理順を示したフローチャートである。

【００１９】また、他にも本文領域末尾の文字の文字種

6

（英数、カタカナ）が、対応する本文領域先頭の文字の文字種と一致しているか否かによっても接続難易度が得られる。さらには、括弧が開いたまま終わっている領域は、閉じ括弧から始まる領域に接続する可能性が大きいように、括弧記号がある場合の位置関係からも接続難易度が得られる。これらの方法により、図１４の文書画像について、見出し領域の前後の文書領域について接続難易度を求めると、図９のようになる。これらの結果から、図１４の文書画像について文書領域間の接続難易度を図示すると、図１０のようになる。図中の実線は難易度＋１を、破線は難易度０を表す。

【００２０】次に、得られた接続難易度を、着目領域をまたぐ方向（横方向）と着目領域に従う方向（縦方向）にそれぞれ累積し、得られた累積値を比較することで接続方向を判別する（Ｓ４２）。すなわち、着目領域をまたぐ方向の接続難易度の累積値が大きければ、着目領域はセパレータとして機能しないものと判別される。次に得られた接続方向に基づき、本文領域間の接続がなされ抽出領域のラベリングを終了する（Ｓ４３）。これらの処理により図１４の文書画像をラベリングした結果が図１１となる。

【００２１】この第２の発明の実施例では、文書段落領域の間に見出しがある場合でも、その前後の文書段落領域の文字を認識して接続難易度を算出することにより、接続方向が確実に判定されて正しくラベリングが行われるようになる。なお、第２の発明の実施例では、言語処理の対象とする部分を段落領域の比較対象となる先頭と末尾に限定することで処理の高速化が可能になる。さらには、図６のフローチャートにおけるＳ３３とＳ３４、Ｓ３４とＳ３６～Ｓ４０は並列処理が可能であるので処理速度をさらに高速にすることも可能である。また、Ｓ３４では全ての本文領域について文字認識をしているが、Ｓ４０で接続の可能性ありとされた本文領域についてのみ文字認識をすればさらに高速化が可能である。

【００２２】次に、第３の発明の実施例について説明する。前述した第２の発明の実施例では、接続難易度を算出する際に単語辞書を参照するためにメモリのアクセス時間が長くなり、高速化に限界がある。そこで第３の発明では、処理速度を増すため、第１の発明および第２の発明でそれぞれ求めた接続難易度を用いて接続方向を判定してラベリングするようにした。つまり、通常は第１の発明により比較的高速で接続難易度を求めて接続方向を判定するが、第１の発明の接続難易度だけでは充分に判定できない場合に第２の発明で求めた接続難易度により接続方向を判定してラベリングするようにしたものである。それにより、確実で高速なラベリングが可能になる。

【００２３】

【発明の効果】以上述べたように第１の発明によれば、見出し段落領域前後に位置する文書段落領域の先頭およ

7

び末尾の空白の有無から文書段落領域間の接続の難易度が算出されて、見出し段落領域前後の文書段落領域の接続方向が判別される。それにより、新聞等の構造が複雑な文書であっても、見出しをセパレータとして誤認識することなく高速にラベリングすることができる。

【００２４】第２の発明によれば、見出し段落領域前後に位置する文書段落領域の先頭および末尾の認識結果を互いに比較して文書段落領域間の接続の難易度を算出することにより、見出し段落領域前後の文書段落領域の接続方向が判別される。それにより、新聞等の構造が複雑な文書であっても、見出しをセパレータとして誤認識することなく正確にラベリングすることができる。

【００２５】第３の発明によれば、第１の発明と第２の発明を組み合わせ、通常は比較的処理速度の速い第１の発明により文書段落領域間の接続難易度を求めて接続方向を判別する。さらに、第１の発明では方向ごとの接続難易度の値に明瞭な差が認められない場合に、第２の発明を用いて文書段落領域の文字を認識して接続難易度を再度求めることにより高精度に接続難易度を求めて接続方向を判別する。その結果、入力文書画像から抽出された文書領域を高速でかつ高精度にラベリングすることが可能になる。

【図面の簡単な説明】

【図１】第１の発明の実施例の処理動作を示すフローチャートである。

【図２】第１の発明の実施例におけるラベリング対象の段落領域の説明図である。

8

【図３】第１の実施例において接続可能な近接領域を検索する説明図である。

【図４】第１の実施例において各段落領域間の接続難易度を示す説明図である。

【図５】第１の実施例において文書画像をラベリングした結果を示す説明図である。

【図６】第２の発明の実施例の処理動作を示すフローチャートである。

【図７】第２の実施例において文書画像から作成された探索木を示す図である。

【図８】第２の実施例の接続難易度計算の処理順を示したフローチャートである。

【図９】第２の実施例において得られた接続難易度の一例を図である。

【図１０】第２の実施例において得られた接続難易度を文書画像に示した図である。

【図１１】第２の実施例において文書画像をラベリングした結果を示す説明図である。

【図１２】従来の処理例を示すフローチャートである。

【図１３】従来のラベリング結果を示した図である。

【図１４】従来のラベリング結果を示した図である。

【符号の説明】

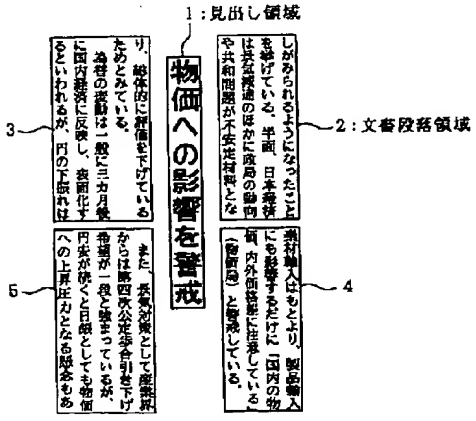
1　見出し領域

2～5　文書段落領域

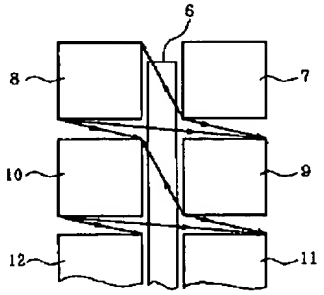
6　見出し領域

7～12　文書段落領域

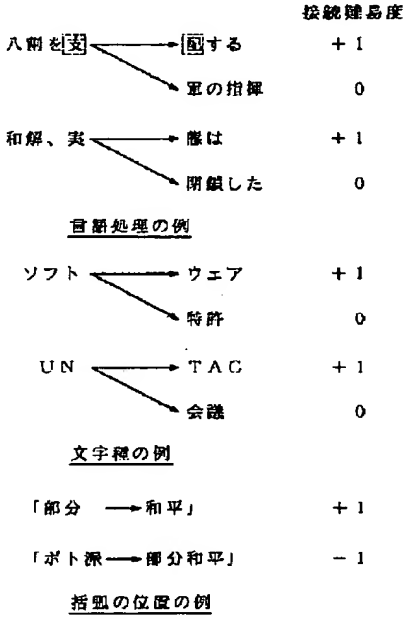
【図２】



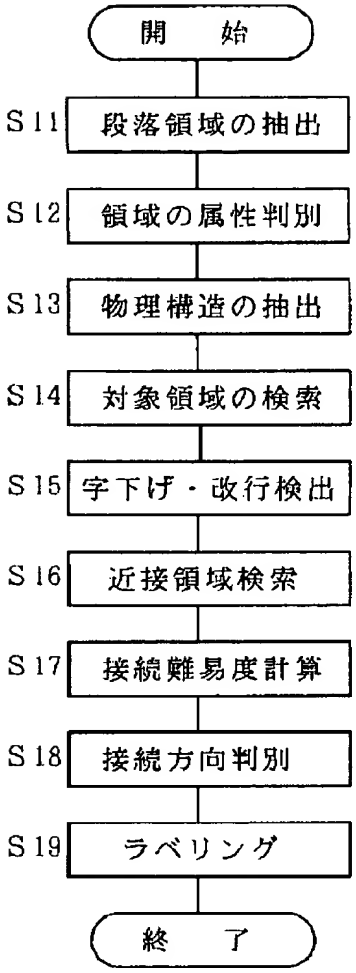
【図３】



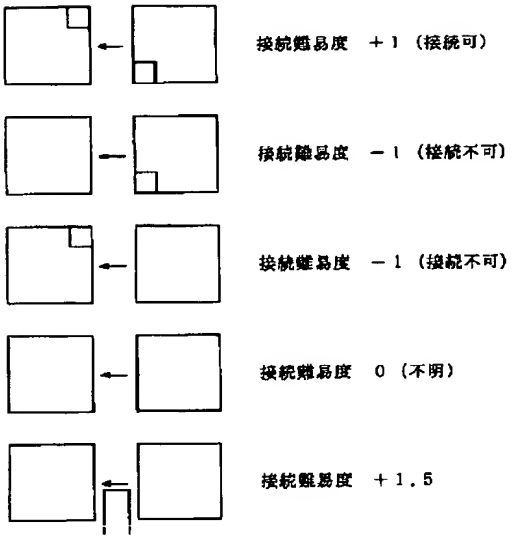
【図９】



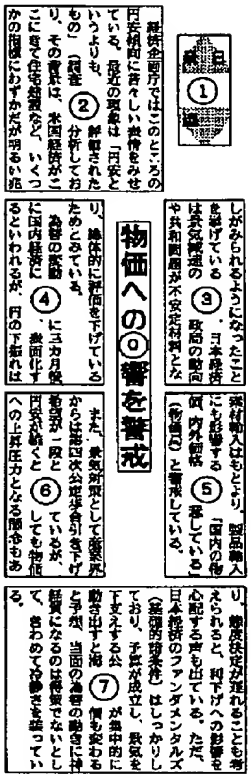
【図1】


開　始
S11
段落領域の抽出
S12
領域の属性判別
S13
物理構造の抽出
S14
対象領域の検索
S15
字下げ・改行検出
S16
近接領域検索
S17
接続難易度計算
S18
接続方向判別
S19
ラベリング
終　了

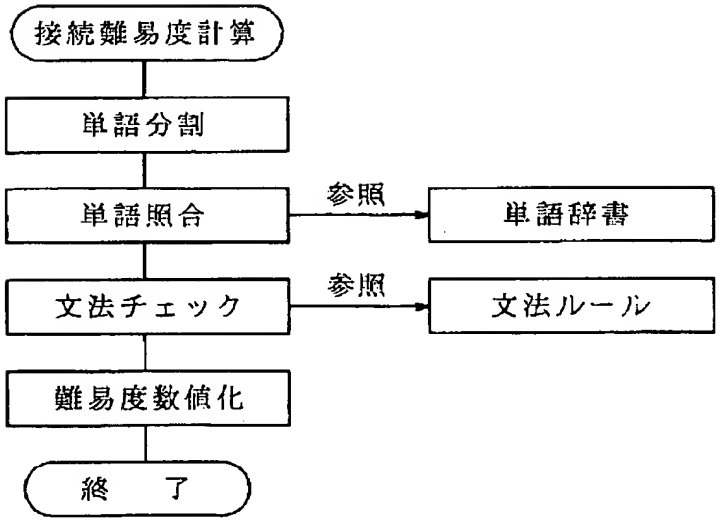
【図4】


接続難易度　＋1（接続可）
接続難易度　－1（接続不可）
接続難易度　－1（接続不可）
接続難易度　0（不明）
接続難易度　＋1.5

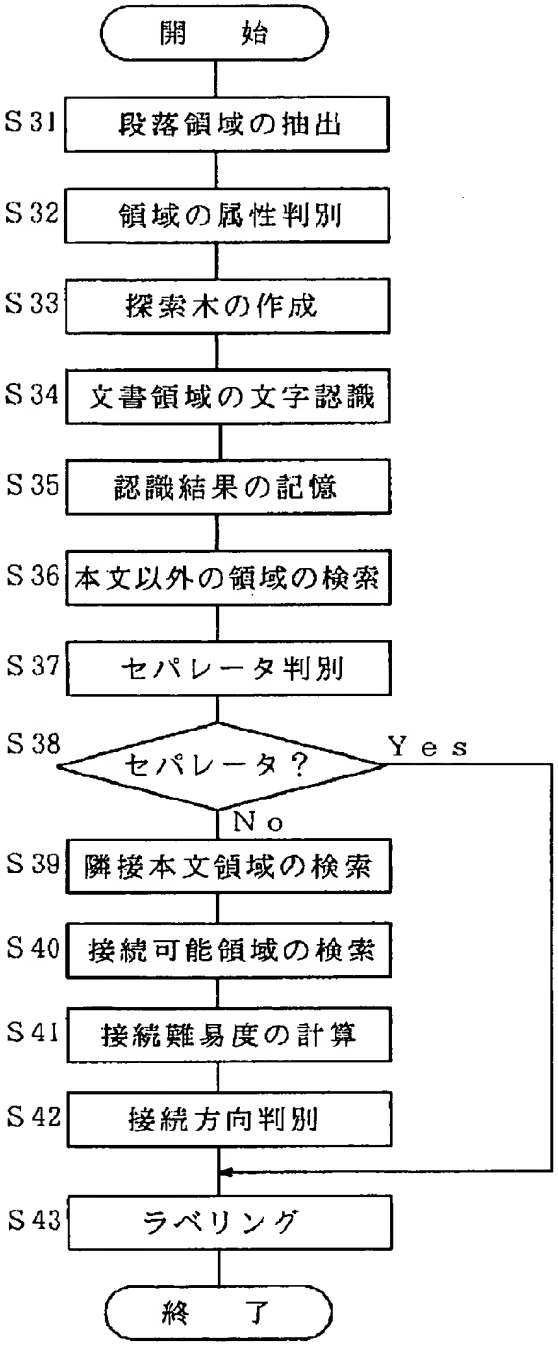
【図5】


物価への①響を警戒

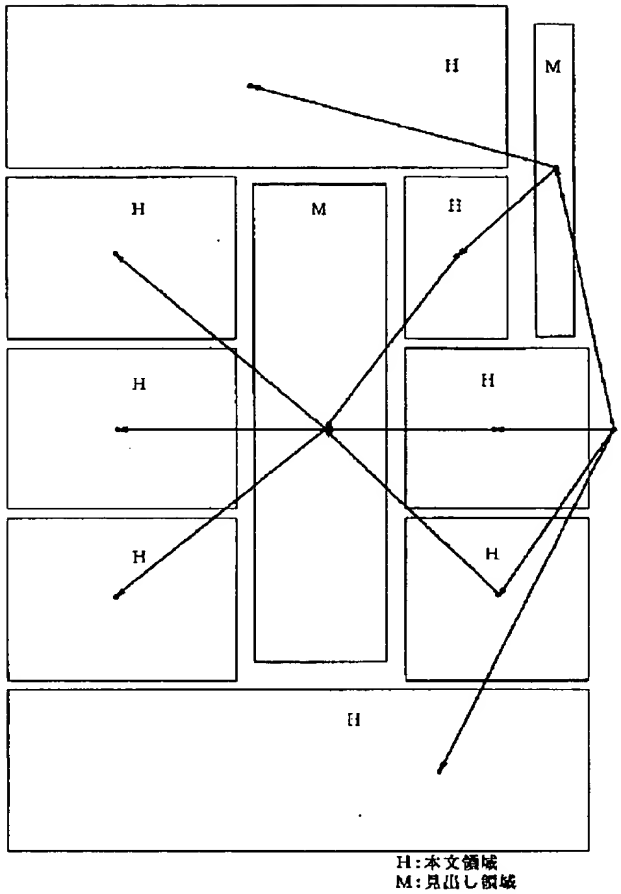
【図8】


接続難易度計算
単語分割
単語照合
参照
単語辞書
文法チェック
参照
文法ルール
難易度数値化
終　了

【図6】


開　始
S31
段落領域の抽出
S32
領域の属性判別
S33
探索木の作成
S34
文書領域の文字認識
S35
認識結果の記憶
S36
本文以外の領域の検索
S37
セパレータ判別
S38
セパレータ？
Yes
No
S39
隣接本文領域の検索
S40
接続可能領域の検索
S41
接続難易度の計算
S42
接続方向判別
S43
ラベリング
終　了

【図7】


H
M
H
M
H
H
H
H
H
H:本文領域
M:見出し領域

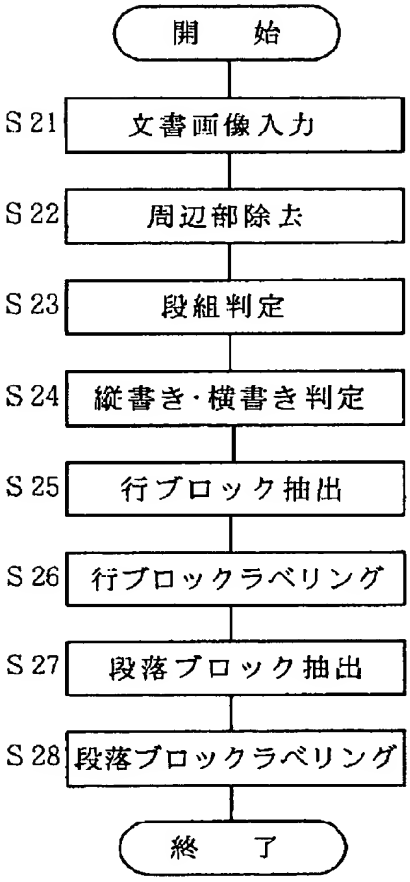
【図10】


カンボジアPKO
難しい日本の対応
選挙妨害で緊迫
難易度＋1
難易度0

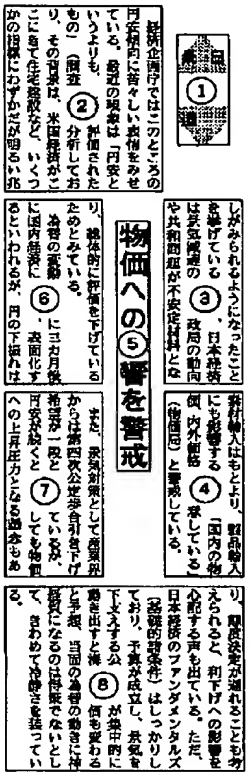
【図11】


カンボジア①/PKO
難しい日本の対応
選挙妨②害で緊迫

【図１２】


開　始
S21
文書画像入力
S22
周辺部除去
S23
段組判定
S24
縦書き・横書き判定
S25
行ブロック抽出
S26
行ブロックラベリング
S27
段落ブロック抽出
S28
段落ブロックラベリング
終　了

【図１３】


物価への⑤警戒

【図１４】